学校编码：10384
学号：23020081153259
分类号_____密级_____
UDC_____

厦门大学

硕士学位论文

# 视觉模式识别中的判别性视觉词选择研究

# Discriminative Visual Words Research on Visual Pattern Recognition

吴迪炜

指导教师姓名：曲延云　副教授

专业名称：计算机应用技术

论文提交时间：2011 年　5　月

论文答辩日期：2011 年　　月

学位授予日期：2011 年　　月

答辩委员会主席：

评　阅　人：

2011 年 5 月

# 厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：吴迪炜

2011 年 6 月 3 日

# 厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：吴迪炜

2011 年 6 月 3 日

# 摘 要

一般目标类识别和行为识别是当前计算机视觉界研究的热点问题。词袋模型为解决该类问题提供了一个基本框架。词袋模型涉及四个关键的因素：局部特征检测、局部特征描述、视觉词典的构造及分类器的设计。传统的视觉词典是由聚类算法得到，把聚类中心作为视觉词。其局限性在于，那些平凡的视觉词，就像文本处理中的冠词一样，大量出现在词典中，使词典庞大，造成图像量化表示的维数高，从而使得计算复杂性高。针对这一局限性，本文重点研究判别性视觉词的提取，在模式分类中提取最具判别性的视觉词，减小词典的规模，降低计算复杂性。另外，传统的词袋模型假设视觉词之间是独立的，然而，视频图像的视觉词之间在时间上是有关联的，本文针对行为识别研究具有时间关联性的判别性视觉词选择。本文主要的研究工作和贡献如下：

1. 针对一般类目标识别，提出基于假设检验的判别性视觉词提取算法。本文分别利用T检验，秩和检验，柯尔莫可洛夫-斯米洛夫检验三种假设检验方法计算各自的最优置信集的视觉词，三种检验得到的所有视觉词构成判别词典。将本文所提算法应用于Caltech、UIUC及Xerox标准图像库，实验结果验证了所提算法的有效性。

2. 针对一般类目标识别，提出基于最大边际多样性的视觉判别词提取方法。对图像的局部特征集，进一步放宽假设检验的限制，在不作任何先验假设的情况下，采用使边际多样性最大化的优化模型，求解判别性视觉词。将本文所提算法应用于Caltech、UIUC及Xerox标准图像库，并与传统的K-means聚类形成的词典比较，实验结果显示，算法在计算时间及分类精度上有了一定的提高。

3.针对行为识别，提出基于格兰杰因果关系检验的判别性视觉词提取算法。将视频中局部特征之间的时间关联性考虑到视觉词的构造中，利用格兰杰因果关系，建立视频视觉词时间上的关联性，同时，结合最大边际多样性对视觉词进行选择。将本文算法应用于KTH人类行为视频库，鼠行为视频库等标准行为识别数据库上进行实验，实验结果验证了算法的有效性。

**关键词**：一般目标类识别；词袋模型；假设检验；边际多样性；格兰杰检验

# Abstract

General objection recognition and action recognition are hotspots in computer vison study presently. The bag-of-words model is providing a general solving framework for these two problems. There are four key factors about the bag-of-words model: local features detection, local features description, visual dictionary construction and classification. Typically, visual dictionary is built by k-means algorithm, and the cluster centers as visual words. Its limitation is that k-means generalizes so many trivia words like article words in text processing which expand the dictionary and induce the high-dimension problem and high complexity. To break out this limitation, we focus on the selection of discriminative visual words so as to extract the most discriminative visual words, reduce the dictionary scale and lower the calculation complexity. The typical bag-of-words model assumes that the visual words are independently distributed, but video visual words have spatiotemporal relationships. We study about selecting discriminative visual words by introducing temporal relation between these words. There are the major researches and contributions:

1. For general object recognition, we propose a discriminative visual words selection algorithm based on hypothesis testing. We adopt three kinds of hypothesis testing methods including T test, rank sum test and Kolmogorov-Smiromov test to calculate the best confidence sets respectively, and then build a discriminative dictionary by uniting the three sets. Experimental result of the Caltech, UIUC and Xerox dataset shows that the algorithm is effective.

2. For general object recognition, we propose another discriminative visual words selection algorithm based on max marginal diversity. By

relaxing constrain on hypothesis testing, we choose max marginal diversity without any prior to optimize the model and select discriminative words. Compared with typical K-means, the proposed algorithm improves performance on complexity and precision.

3. For action recognition, we propose a discriminative visual words selection algorithm based on Granger test. We design a method of choosing the temporal correlative words by considering the temporal correlationship between visual words, adopting Granger test to build the time relationship between words and combining the max marginal diversity. The algorithm is applied to KTH human action dataset and mouse behavior dataset. Experimental results validate its effectiveness.

Key words: general object recognition; bag-of-words model; hypothesis testing; max marginal diversity; Granger test.

# 目 录


**第一章 绪论 ..... 1**

**1.1 研究意义 ..... 1**

**1.2 研究现状 ..... 2**

**1.3 本文的主要工作及内容安排 ..... 5**

**第二章 词袋模型 ..... 8**

**2.1 局部特征检测 ..... 9**

**2.2 局部特征描述 ..... 10**

**2.3 视觉词典构造 ..... 10**

**2.4 分类器设计 ..... 11**

2.4.1 线性可分的最优分类面 ..... 11

2.4.2 线性不可分的最优分类面 ..... 13

2.4.3 SVM 的核函数 ..... 14

2.4.4 SVM 多分类器算法 ..... 14

**2.5 小结 ..... 15**

**第三章 基于假设检验的视觉判别词选择 ..... 16**

**3.1 目标类的特征检测与描述 ..... 16**

3.1.1 Canny 边缘检测 ..... 16

3.1.2 基于 SIFT 的特征描述 ..... 17

**3.2 基于假设检验的视觉判别词选择 ..... 21**

3.2.1 参数检验与非参数检验 ..... 22

3.2.2 T 检验 ..... 22

3.2.3 秩和检验 ..... 23

3.2.4 K-S 检验 ..... 24

3.2.5 三种检验工具与判别性视觉词筛选 ..... 24

**3.3 本节实验 ..... 25**

3.3.1 Caltech 及 UIUC5 类图像库 ..... 25

3.3.2 Xeror7 图像库 ..... 26

3.3.3 本节测试方式与评价标准 ..... 26

3.3.4 Caltech&UIUC5 图像库实验结果 ..... 27
3.3.5 Xerox7 图像库实验结果 ..... 30
3.3.6 实验图库测试时间 ..... 33
**3.4 小结 ..... 33**
**第四章 基于最大边际多样性视觉判别词选择 ..... 34**
**4.1 最大边际多样性的判别词筛选策略 ..... 34**
**4.2 本节实验 ..... 35**
4.2.1 Caltech&UIUC5 图像库实验结果 ..... 35
4.2.2 Xerox7 图像库实验结果 ..... 37
4.2.3 实验图库测试时间 ..... 39
**4.3 小结 ..... 40**
**第五章 基于格兰杰因果判别词选择的行为识别 ..... 41**
**5.1 行为的特征检测与描述 ..... 41**
5.1.1 P. Dollár 周期检测 ..... 41
5.1.2 主成分分析 ..... 42
5.1.3 光流法表示行为的运动信息 ..... 43
5.1.4 行为的特征描述 ..... 44
**5.2 对视频片段的格兰杰检验 ..... 46**
5.2.1 自回归模型 ..... 46
5.2.2 格兰杰因果关系检验 ..... 48
5.2.3 定阶问题与 AIC 准则 ..... 50
5.2.4 格兰杰检验与视觉词筛选 ..... 51
**5.3 本节实验 ..... 52**
5.3.1 KTH 人类行为数据库 ..... 52
5.3.2 鼠行为数据库 ..... 53
5.3.3 测试方式与评价标准 ..... 54
5.3.4 KTH 人类行为视频库实验结果 ..... 55
5.3.5 鼠行为视频库实验结果 ..... 57
5.3.6 实验视频库测试时间 ..... 60
**5.4 小结 ..... 60**
**第六章 全文总结及展望 ..... 62**
**6.1 总结 ..... 62**

6.2 展望 ..... 62
参考文献 ..... 64
研究生期间参加的科研活动及科研成果 ..... 70
致谢 ..... 71

# Contents

**Chapter 1 Introduction.....1**

**1.1 Research significance.....1**

**1.2 Research status.....2**

**1.3 Contributions and Outline.....5**

**Chapter 2 Bag-of-words model.....8**

**2.1 Local feature detectors.....9**

**2.2 Local feature descriptors.....10**

**2.3 Construction of visualdictionary.....10**

**2.4 Classifiers.....11**

2.4.1 Best linear separable classify plane.....11

2.4.2 Best linearly nonseparable classify plane.....13

2.4.3 SVM kernel function.....14

2.4.4 SVM multiclass classification algorithm.....14

**2.5 Conclusion.....15**

**Chapter 3 Visual words selection based on hypothesis testing.....16**

**3.1 Feature detection and description for objects.....16**

3.1.1 Canny edge detecor.....16

3.1.2 SIFT descriptor.....17

**3.2 Visual words selection based on hypothesis testing.....21**

3.2.1 Parametric test and nonparametric test.....22

3.2.2 T test.....22

3.2.3 Rank-sum test.....23

3.2.4 K-S test.....24

3.2.5 Hypothesis testing methods for words selection.....24

**3.3 Experiments.....25**

3.3.1 Caltech & UIUC 5-classes dataset.....25

3.3.2 Xeror7 dataset.....26

3.3.3 Testing methods and evaluation criteria.....26

3.3.4 Caltech & UIUC 5-classes dataset results.....27
3.3.5 Xeror7 dataset results.....30
3.3.6 Testing time analysis.....33
**3.4 Conclusion.....33**
**Chapter 4 Visual words selection based on MMd.....34**
**4.1 maximum marginal diversity.....34**
**4.2 Experiments.....35**
4.2.1 Caltech & UIUC 5-classes dataset results.....35
4.2.2 Xeror7 dataset results.....37
4.2.3 Testing time analysis.....39
**4.3 Conclusion.....40**
**Chapter 5 Action recognition based on Granger test.....41**
**5.1 Feature detection and description for actions.....41**
5.1.1 P. Dollár periodic detector.....41
5.1.2 Principal Component Analysis.....42
5.1.3 Optical flow method.....43
5.1.4 Action descriptor.....44
**5.2 Granger test to vedio clips.....46**
5.2.1 Autoregression model.....46
5.2.2 Granger test.....48
5.2.3 Order determination & Akaike Information Criterion.....50
5.2.4 MMd+Granger algorithm.....51
**5.3 Experiments.....52**
5.3.1 KTH humana action dataset.....52
5.3.2 mouse action dataset.....53
5.3.3 Testing methods and evaluation criteria.....54
5.3.4 KTH humana action dataset results.....55
5.3.5 mouse action dataset results.....57
5.3.6 Testing time analysis.....60
**5.4 Conclusion.....60**
**Chapter 6 Summary and future work.....62**
**6.1 Summary.....62**

6.2 Future work ..... 62
**References** ..... **64**
**Publications** ..... **70**
**Acknowledgement** ..... **71**

# 第一章 绪论

## 1.1 研究意义

基于视觉的模式识别一直是计算机视觉研究领域中的挑战性问题。尽管计算机视觉发展了几十年，但机器自动识别和分析的能力还远远落后于人类的视觉。神经学领域和生物学领域及其它相关领域的学者们都投入极大的兴趣来研究人类视觉系统的机理，但到现在为止，人们也尚未完全搞清楚人类视觉机理。这就造成当前无法完全模拟人类视觉系统。另外，现有的计算机的构造也与人类的脑神经系统结构不同，其计算方式无法模拟人类脑神经系统。人类视觉系统的优越性是长期进化的结果，人类视觉能识别超过 30000 类的目标，不仅如此，在多种复杂的环境中，人类视觉对目标的识别还具有鲁棒性，比如，对颜色具有识别的恒常性，即使是强光下的绿色苹果，其部分地方反光泛白，人类也知道是绿色的苹果。人类视觉的识别能力不仅体现在对静态视觉模式的识别上，更体现在对动作、行为、事件的分析和理解上。就静态视觉模式识别来说，自然场景中的目标往往受光照、遮挡、尺度变换以及拍摄视角影响等，如图 1.1 所示，这就造成同类目标差异性较大，甚至有可能接近异类目标的外观。对机器视觉来说，要成功识别多类目标依然极具挑战性。一个更加令机器视觉望尘莫及的任务是对动作和事件的分析和理解。人类往往通过时间序列的视场内容，判断出发生了什么事情，并能激发出人类特有的感情，如图 1.2 所示，是一个人类见面握手的事件。而对机器视觉来说，对事件、行为的识别、分析和理解则刚刚起步。


**图 1.1：**自然场景中的目标

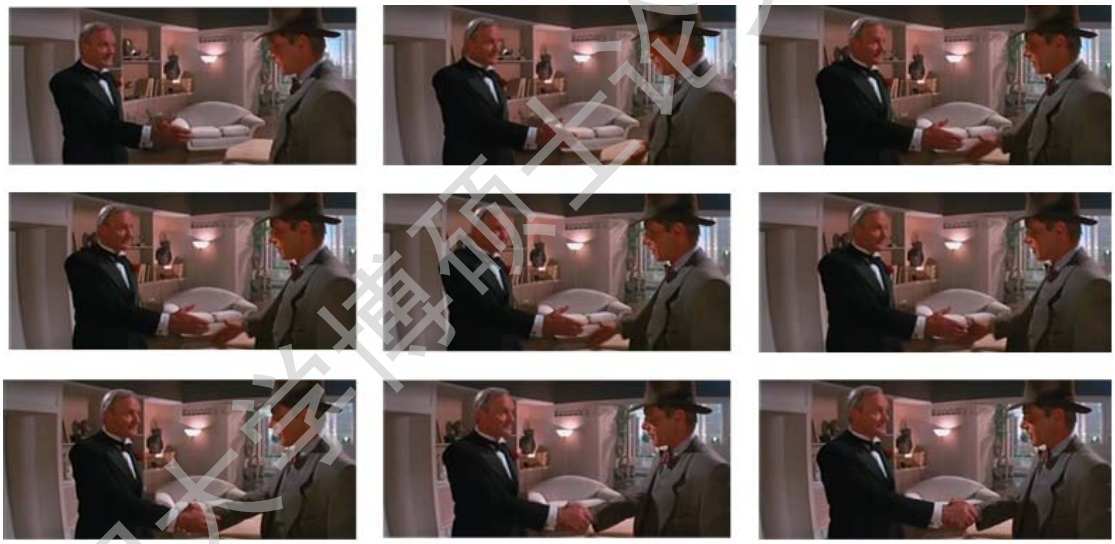

**图 1.2：**人类见面握手

## 1.2 研究现状

众多的学者在视觉模式识别上做了大量的工作，取得了阶段性的进展。其中，有两个关键的因素影响着该类问题的解决，一是图像的表示，二是分类器的设计。在一般目标类的图像识别中，流行的图像表示之一是基于词袋模型[1]的图像表示方法，它为多类图像及视频图像序列提供了一个图像表示的一般性框架。流行的词袋模型包括四个关键的步骤：

（1）局部特征检测；

（2）局部特征描述；

（3）视觉词典构造；

（4）分类器设计。

词袋模型把图像或视频的信息映射到视觉词的集合，简单，高效并且对仿射变换，遮挡，光照和类内变化的具有鲁棒性，既保存了局部特征又有效地压缩原数据，使得自然语言处理的各项技术和方法在计算机视觉领域得到有效地应用。

静态目标识别方面，Li FeiFei[1][2]和 Sivic[3]基于词袋模型提取自然场景中的主题，Lazebnik[4]，Leung, T.[5]，Varma, M.[6]等人在纹理分类（texture categorization）中采用词袋模型也获得理想的实验结果；视频处理方面，P. Dollár等人[7]基于词袋模型对实验室场景下的人类表情、行为以及老鼠行为进行分析，Karthir Prabhakar[8]在词袋模型的框架下对生活社交场景的视频进行关联信息的数据挖掘，并利用该信息进行行为分类，Guangyu Zhu[9]在广播电视的网球运动视频上利用词袋模型进行行为识别，词袋模型在各类不同情境下的行为识别实验中都有着上佳表现。在词袋模型的框架下，一般类目标识别和行为识别的解决方法可以分为两大类，其一是直接采用词袋模型而不考虑局部特征之间的空间关系或时间关系，其二是将词袋模型与时空关系结合在一起。

在上文提及的研究工作中，分类能力的高低取决于图像或者视频的表示方法和分类器的设计方法。对于如何表示一幅图像或者一段视频，我们需要处理三个方面的问题：第一，采样，即如何获得关键点或者关键区域；第二，描述，即如何表示关键点或区域；第三，如何定量分析局部特征的连续空间。对于本文中关注的视觉词选择问题，研究者们提出了许多改善方法。我们可以按照这些改进方法的与分类器设计的关系将其分成两类。

（1）结合分类器的词典构建

研究者将改进方法与分类器相关联，词典的生成过程伴随着分类器的训练过程。视觉词选择算法本身作为组成部分嵌入到分类器中，能够选出更适合某个分类器的特征，但是复杂度较高，执行速度慢。

在静态目标识别方面：Farquhar[10]等人使用高斯混合模型对每一类关键点

的密度函数进行建模，Moosmann[11]等人使用随机聚类森林来构建判别性视觉词典，Perronnin[12]通过联合建立全局词典与专门词典的方式来构建自适应词典， Larlus 和 Jurie[13]应用多元高斯混合 LDA 模型来构建词典，Liu Yang[14]等人针对具体类别选择不同类别专有词。

在视频处理方面：Juan Carlos Niebles[15]等人应用概率潜在语义分析（PLSA）模型来构建词典，Lixin Duan[16]将视频片段按照时空尺度的不同进行多次分割，并采用自适应多核学习技术对生成的分类器进行赋予权重，生成多层次的词典，Adriana Kovashka[17]提出构造包含近邻位置及其方向信息的多层次判别性时空特征，并使用多核学习技术调整时空特征的权重，构造判别性词典。

（2）基于数据特征的词典构建

另外一类研究选择与分类器分离的改进方法，根据数据集本身固有的特性进行视觉词选择，选词方法独立于后续使用的分类器，计算效率较高，适用于规模较大的应用，缺点是分类精度往往低于前一种方法。

在静态目标识别方面：Jurie 和 Triggs[18]提出了一个基于互信息的选词策略，Moulin[19]等人提出组合 TF-IDF（term frequency–inverse document frequency）权重的构建多词典的选词方法，Winn[20]等人对滤波器组的响应进行聚类来生成一个精简的视觉词典，Lazbnik[21]使用半局部部分（semi-local parts）来表示一个图像。

在视频处理方面：Guangyu Zhu[9]采用一组光流梯度直方图作为运动视频特征描述，在此基础上利用词袋模型进行行为识别，Olivier Duchenne [22]等人通过最大边际聚类来构造词典，Karthir Prabhakar[8]等人通过抽取视频中有时间因果关系的因果集，手动提取有判别性的聚类中心子集来构造词典，从而提高真实场景中社交游戏的分类精度。图 1.3 给出图像的判别性词示例，图上半部分为原始图片，下半部分为判别性词所在区域，图 1.4 为视频的判别性词示例，其中绿色圆圈部分为判别性词的所在区域。